https://doi.org/10.5194/egusphere-2024-361
Preprint. Discussion started: 19 February 2024

Unveiling the optimal regression model for source apportionment of the oxidative potential of PM

Vy Dinh Ngoc Thuy[1], Jean-Luc Jaffrezo[1], Ian Hough[1], Pamela A. Dominutti[1], Guillaume Salque Moreton[2], Grégory Gilles[3], Florie Francony[4], Arabelle Patron-Anquez[5], Olivier Favez[6,7], Gaëlle Uzu[1]

[1] Université Grenoble Alpes, CNRS, IRD, INP-G, INRAE, IGE (UMR 5001), F-38000 Grenoble, France

[2] Atmo AuRA, 69500 Bron, France

[3] Atmo Sud, 13294 Marseille, France

[4] Atmo Nouvelle Aquitaine, 33692 Merignac, France

[5] Atmo Hauts de France, 59044 Lille, France

[6] INERIS, Parc Technologique Alata, BP 2, 60550 Verneuil-en-Halatte, France

[7] Laboratoire central de surveillance de la qualité de l'air (LCSQA), 60550 Verneuil-en-Halatte, France

Correspondance to: gaelle.uzu@ird.fr

Abstract

The capacity of particulate matter (PM) to generate reactive oxygen species (ROS) in vivo leading to oxidative stress, is thought to be a main pathway for the health effect of PM inhalation. Exogenous ROS from PM can be assessed by acellular oxidative potential (OP) measurements as a proxy of the induction of oxidative stress in the lungs. Here, we investigate the importance of OP apportionment methods on OP repartition by PM sources in different types of environments. PM sources derived from receptor models (e.g. EPA PMF) are coupled with regression models expressing the associations between PM sources and OP measured by ascorbic acid (OP_{AA}) and dithiothreitol assay (OP_{DTT}). These relationships are compared for eight regression techniques: Ordinary Least Squares, Weighted Least Squares, Positive Least Squares, Ridge, Lasso, Generalized Linear Model, Random Forest, and Multilayer Perceptron. The models are evaluated on one year of PM_{10} samples and chemical analyses at each of six sites of different typologies in France to assess the possible impact of PM source variability on OP apportionment. Source-specific OP_{DTT} and OP_{AA} and out-of-sample apportionment accuracy vary substantially by model, highlighting the importance of model selection depending on the datasets. Recommendations for the selection of the most accurate model are provided, encompassing considerations such as multicollinearity and homoscedasticity.

Key words: Oxidative potential, source apportionment, OP apportionment.

1. Introduction

Ambient particulate matter (PM) is one of the key contributors to atmospheric pollution and is responsible for approximately 7 million premature deaths worldwide yearly (WHO, 2021). Many epidemiological studies have linked PM exposure to adverse health effects including (i) acute effects studies using time series and related studies to evaluate the immediate impact of PM exposure (Bell et al., 2004; Dominici, 2004; Peng et al., 2009; Pope & Dockery, 2006) and (ii) cohort studies aiming to evaluate the long-term effects of chronic PM exposure (Ayres et al., 2008; Beelen et al., 2014; Crouse et al., 2012, 2015; Pelucchi et al., 2009; Yu et al., 2021). These studies mainly focused on the association with PM mass concentrations. However, various research shows that the impacts of PM also depend on other factors such as chemical composition, size distribution, particle morphology, and biological mechanisms (Crouse et al., 2012). PM's capacity to generate reactive oxygen species (ROS) in vivo has recently been introduced as a pivotal indicator of PM biological mechanism, with direct implications for oxidative stress and cellular damage (Akhtar et al., 2010; Ayres et al., 2008; Leni et al., 2020; Li et al., 2008; Lodovici & Bigagli, 2011; Mudway et al., 2020; Nelin et al., 2012; Rao et al., 2018). The quantification of the PM capacity to oxidize a biological media is called oxidative potential (OP). Various acellular assays of OP have been introduced, differentiating ROS generation mechanisms of PM (Calas et al., 2018; Dominutti et al., 2023). Dithiothreitol (DTT) and ascorbic acid (AA) assays are two of the commonly used ones in the literature (Liu & Ng, 2023).

The relationship between PM chemical components and OP activities may identify which components are most prone to generate ROS (Calas et al., 2018, 2019; Crobeddu et al., 2017; Godri et al., 2011; Janssen et al., 2014; Szigeti et al., 2015, 2016; Yang et al., 2014). However, this research pathway struggles with the co-variation between measured and unmeasured PM components (Calas et al., 2018; Weber et al., 2018). An alternative approach is to examine the association between OP and sources of PM obtained using receptor models such as chemical mass balance, positive matrix factorization (PMF), or principal components analysis. PMF is the most popular method for its ability to quantify PM source contributions without extensive prior information on specific sources at the site studied (Belis et al., 2013; Brown et al., 2015; Paatero & Hopke, 2009; Paatero & Tappert, 1994; Viana et al., 2008).

Regression analysis is the most common and effective way to estimate the redox activity of receptor model-derived PM sources. Generally, this is achieved by regression analyses to characterize the relationship between OP activities (nmol min^{-1} m^{-3}) and PM sources contribution (µg m^{-3}). This approach provides the OP activities attributed to each microgram of each source (nmol min^{-1} µg^{-1}), denoted as intrinsic OP, which can be used to calculate the contribution of each source for each observation day. Numerous regression models can be used for such OP source apportionment (SA), with multiple linear regression fitted by ordinary least squares being the most common regression technique (Bates et al., 2015; Deng et al., 2022; Li et al., 2023; Liu et al., 2018; Shangguan et al., 2022; Verma et al., 2014; Wang et al., 2020; Yu et al., 2019). Further, some studies exclude sources with negative intrinsic OP, assuming that negative OP activities are geochemically nonsensical (Bates et al., 2018; Weber et al., 2018). Additionally, weighted least square can be used to introduce a weighting term, usually using the OP analysis uncertainties to take into account the measurement uncertainties of the OP assays (Borlaza et al., 2021; Daellenbach et al., 2020; Dominutti et al., 2023; Fadel et al., 2023; in 't Veld et al., 2023b; Weber et al., 2021). Finally, non-linear models, such as multilayer perceptron, have been used to try to capture possible non-linearities between OP activities and PM sources (Borlaza et al., 2021; Elangasinghe et al., 2014; Wang et al., 2023). However, no study to date has compared the performance and applicability of these various regression models. Each model implies different assumptions which should be carefully considered when selecting a given model.

This study aims to evaluate the variability in OP SA techniques by comparing eight regression techniques: multiple linear regression fitted by ordinary least squares (OLS), weighted least squares (WLS), positive least squares (PLS), Ridge regression (Ridge), Least Absolute Shrinkage and Selection Operator (Lasso), generalized linear model (GLM), random forest (RF), and multilayer perceptron (MLP). These techniques are applied to apportion OP_{AA} and OP_{DTT} to PM sources at six sites in France. The PM SA outputs have been published previously in Weber et al., (2021), using a harmonized PMF methodology based on one year of sampling with similar chemical analyses for a large set of chemical tracers. The results of the OP SA models are compared with regard to the estimated intrinsic OP of each source, the out-of-sample accuracy of the apportionment, and the assumptions inherent in each model. The most appropriate model at each site is compared with OLS to quantify the difference between choosing a model based on data characteristics vs. using the most common approach. Finally, this study provides guidelines for selecting the most suitable model in the strategy for OP contribution regarding sources of PM. This holds particular significance in the context of the implementation of OP monitoring as a novel air quality metric as foreseen in research programs (such RI-Urbans) and in the process of the revision of the European Directive 2008/50/CE.

2. Methodology

2.1. General organisation of this work

Figure 1 illustrates the general workflow of this work. Sections 2.2, 2.3, and 2.4 describe the methods used to analyse the temporal evolution of PM sources and OP activities, identify collinearity among PM sources, and examine homoscedasticity in the relationship between OP activities and PM sources. Section 2.5 describes the eight regression techniques (OLS, WLS, PLS, Ridge, Lasso, GLM, RF, and MLP), used for OP SA. Each technique is applied to each site separately using OP_v (nmol min^{-1} m^{-3}) as the dependent variable and PM sources (µg m^{-3}) as independent variables. The coefficient of the regression called the intrinsic OP of the source (nmol min^{-1} $µg^{-1}$), represents the capacity of each µg of PM from the given source to generate oxidative stress; the higher the intrinsic OP of a source, the more redox-active. Each model is trained on a randomly selected (without replacement) 80% subsample of the dataset and validated on the remaining 20%. This process is repeated 500 times to estimate uncertainty, a method particularly needed for sources with strong seasonality. For WLS, PLS, Ridge, and Lasso models, OP analytical errors were used as a weighting, implying that the OP with the high analysis uncertainties has less influence on the model. Section 2.6 describes the statistical validation of the models using root mean square error (RMSE), mean absolute error (MAE), R-square (R^2). The geochemical validation is based on the regression coefficient (the intrinsic OP) of each source. These are calculated separately for the training and testing data and averaged across the 500 sampling iterations.

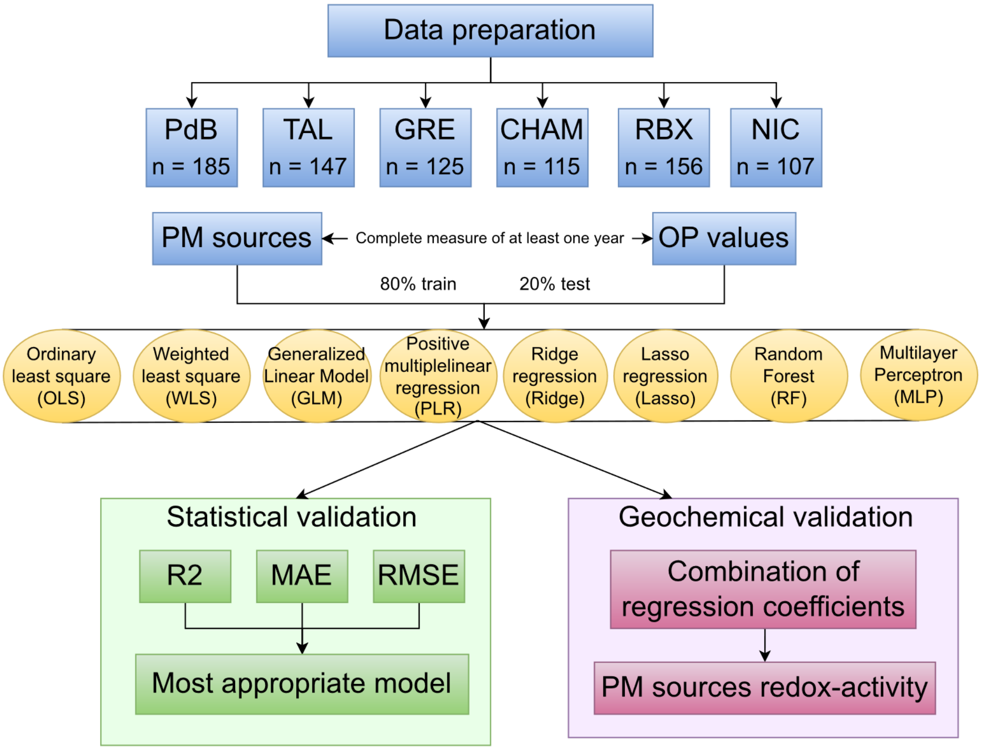


Figure 1. Workflow of the comparison of OP sources apportionment methodology

2.2. Study sites and PM sources

Six French sites are selected in this work for their different typologies: Roubaix and Nice (traffic sites within urban areas), Port-de-Bouc (industrial hotspot), Talence (urban background site), Grenoble and Chamonix (urban background sites in Alpine Valley). At each site, sampling was conducted over at least one year to capture the complete annual evolution of PM and its components. These sites and sampling series were previously used and described by Weber et al. (2019).

In brief, daily filter samples were collected on pre-heated Pallflex quartz fibre filters every third day through high-volume sampling (DA80, Digitel). These filters were analyzed to determine PM's chemical species and OP activities. Further details regarding the chemical species and OP analyses methodology can be found in Weber et al. (2019, 2021). Briefly, the elemental carbon (EC) and organic carbon (OC) were analyzed using the EUSAAR2 thermo-optical protocol with a Sunset Lab analyser. Major ionic components (Cl-, NO_3^-, SO_4^{2-}, NH_4^+, Na^+, K^+, Mg^{2+}, Ca^{2+}) and methanesulfonic acid (MSA) were measured by ion chromatography (IC). Anhydro-sugars and saccharides (including levoglucosan, mannosan, arabitol, sorbitol, and mannitol) were analysed by high-performance liquid chromatography with pulsed amperometry detection (HPLC-PAD). Major and trace elements (Al, Ca, Fe, K, As, Ba, Cd, Co, Cu, La, Mn, Mo, Ni, Pb, Rb, Sb, Sr, V, and Zn) were determined by inductively coupled plasma atomic emission spectroscopy or mass spectrometry (ICP-AES or ICP-MS). Furthermore, colocated PM10 measurements were conducted automatically at each site using the Tapered Element Oscillating Microbalance equipped with a Filter Dynamics Measurement System (TEOM-FDMS).

We used the PM sources identified by Weber et al., (2019), who performed a separate PMF for each site using a harmonized approach for all sites (same chemical species and measurement methods, same procedure to estimate uncertainties, same constraints on the preliminary solutions). Table 1 provides a data description, including the sampling duration, the number of samples collected, and the identified PM sources at each site, while Figure 2 presents the localisation of the sites in France, together with the respective proportion of each PM source at each site.

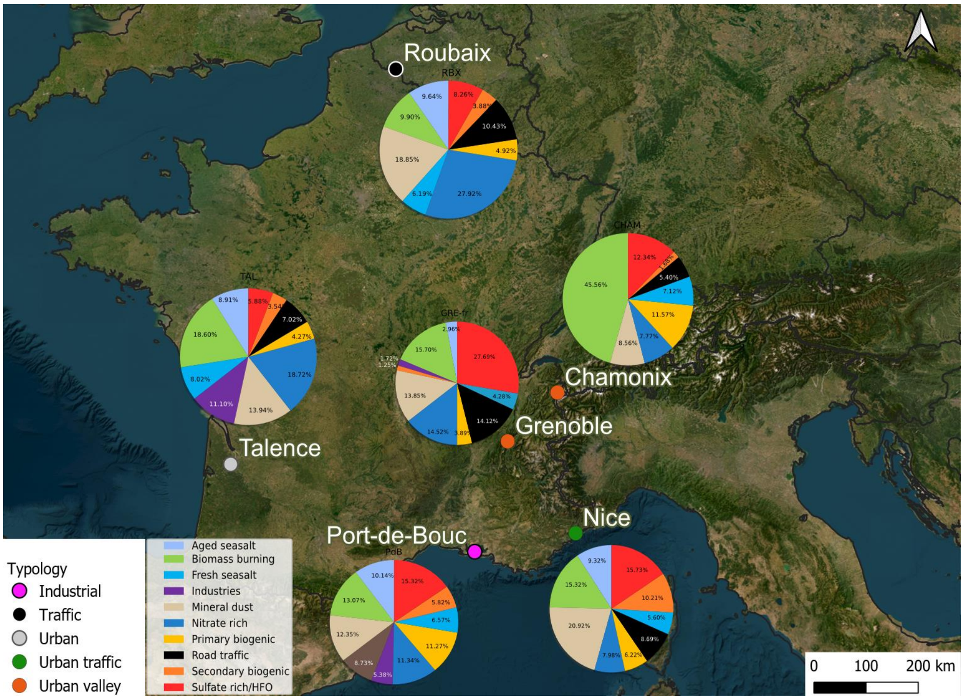


Figure 2. The location of the selected sites for this study. The small colored dots represent the typology of sites. The pie charts are the PM10 source apportionment for each site with the colors identifying the PM sources. Background photography from ESRI satellite.

Table 1. Data description

	PdB	**TAL**	**GRE-fr**	**CHAM**	**RBX**	**NIC**
Name	Port de Bouc	Talence	Grenoble	Chamonix	Roubaix	Nice
N of samples	185	147	125	115	156	107
Sampling dates	2014-06 to 2016-06	2012-02 to 2013-04	2017-02 to 2018-03	2013-11 to 2014-10	2013-01 to 2014-05	2014-07 to 2015-05
N of sources	10	10	10	8	9	9

2.3. OP analysis

OP assays were performed on PM extracted from the filters using simulated lung fluid, as detailed in Calas et al. (2017, 2018). The AA assay involved ascorbic acid, a natural antioxidant in the lungs inhibiting lipid and protein oxidation in the lining fluid, using the method presented by Kelly & Mudway (2003) and further described by Calas et al., (2018). Conversely, the DTT assay used dithiothreitol (DTT) as a chemical surrogate for cellular reducing agents, specifically nicotinamide adenine dinucleotide and nicotinamide adenine dinucleotide phosphate oxidase, thereby replicating in vivo interactions between PM and biological oxidants (Calas et al., 2018; Cho et al., 2005). Both assays measured the consumption of AA or DTT during the assay, i.e., the rate of the transfer of electrons from AA or DTT to oxygen. The assays were conducted with 96-well plates of UV-transparent quality (CELLSTAR, Greiner-Bio), and absorption measurements were acquired using a TECAN spectrophotometer, Infinite M200 Pro, at the wavelengths of 265nm for the AA assay and 412nm for the DTT assay (Calas et al., 2017, 2018, 2019). Each sample extraction was subjected to four analyses; the OP activities in this study represent the mean and the analysis uncertainty is the standard deviation of these four OP analyses. After analysis, the OP activities of each sample were blank-subtracted using lab and field blanks, and normalized using the air sampling volumes and the mass concentration. The resulting OP_V represents the OP activities due to PM per cubic meter of air (nmol min^{-1} m^{-3}).

2.4. Collinearity and heteroscedasticity tests

The result of a regression model strongly depends on the characteristics of the dataset because each model makes assumptions about the data. Two critical assumptions in OLS regression analysis are that (1) there is little collinearity between independent variables (the PM sources in this study), and (2) the variance of the regression residuals is constant (called homoscedasticity). These assumptions should be tested in different ways.

2.4.1. Collinearity

Collinearity occurs when one or more of the independent variables is close to a linear combination of the other independent variables. When collinearity is present, small changes in the data can cause large changes in estimated coefficients, and the estimated standard errors of the coefficients are large. Variance Inflation Factor (VIF) is an indicator of the collinearity between the independent variables (Craney & Surles, 2002; O'Brien, 2007; Rosenblad, 2011). VIF of a specific source is calculated as:

$$VIF_i = \frac{1}{1 - R_i^2}, i = 1, \ldots, p-1 \; (Eq1)$$

https://doi.org/10.5194/egusphere-2024-361
Preprint. Discussion started: 19 February 2024

In this equation, *p* is the number of PM sources, R^2 is the coefficient of determination of a multiple linear regression model between the i^th^ source and the other sources. VIF values of a PM source present a range between 1, and ∞. The higher the VIF values, the greater the collinearity between this PM source and the other ones. A VIF value between 5 and 10 is commonly interpreted as moderate collinearity, while values greater than 10 indicate high collinearity (Craney & Surles, 2002).

2.4.2. Heteroscedasticity

Heteroscedasticity occurs when the variance of regression residuals is not constant but varies for different values of the dependent variable. In this case, the estimated standard errors of the regression coefficients are not reliable. The Goldfeld–Quandt test was developed by Goldfeld & Quandt (1965) to evaluate residual variance in a regression model. To implement the Goldfeld–Quandt test, an OLS regression was performed between OP and PM sources to identify the residual of OP prediction. Next, the PM sources and residual corresponding are divided into three segments: the upper segment is the group with higher PM sources concentration, the lower segment is the group with lower PM sources concentration, and the middle segment, constituting 10% of the moderate PM concentration, is excluded. A subsequent regression analysis is then conducted on the two remaining subgroups to determine the ratio of residual sums of squares. Finally, an F-test is conducted on this ratio to assess whether the variances are the same, with a p-value below 0.05 interpreted as evidence of heteroscedasticity.

The Variance Inflation Factor (VIF) and the Goldfeld–Quandt test were performed in Python 3.9, using the statsmodels 0.14.0 package (Seabold & Perktold, 2010).

2.5. Regression models

The fundamental principle of regression models in this study is to use the PM sources to predict OP activities by identifying the parameters (coefficients and residuals) that minimize an error term (Hastie, 2009). A simple regression model can be represented by Eq. 2, which defines the estimated function of the regression model, and Eq. 3, which estimates the residuals.

$$\hat{y} = f(X) + e \ (Eq2)$$

$$e = y - \hat{y}(Eq3)$$

Here, $\hat{y}$ is the estimated OP (nmol min^{-1} m^{-3}), X are the PM source contributions (µg m^{-3}), y is the observed OP (nmol min^{-1} m^{-3}), and e denotes the residuals (nmol min^{-1} m^{-3}). Each model has certain assumptions and a minimization term, as presented below.

Ordinary least squares (OLS):

OLS is a linear regression technique that minimizes the residual sum of squares. This model is based on several assumptions: (1) **Linearity:** The relationship between OP and PM sources is linear. (2) **Independence:** The PM sources must be independent, with no collinearity. (3) **Homoscedasticity:** The variance of residuals is constant across all values of PM sources. (4) **Normality:** The residuals are normally distributed. In the OLS model, the estimated equation and objective to minimize are defined as follows:

$$\hat{y} = \beta_0 + \sum_{1}^{p} \beta_i * x_i \ (Eq4)$$

$$Minimize: \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \ (Eq5)$$

Here, the β_0 denotes the intercept (nmol min^{-1} m^{-3}), β_i represents the regression coefficient (intrinsic OP, nmol min^{-1} µg^{-1}) of source i, x_i is the concentration of source i (µg m^{-3}), p is the number of PM sources, and m is the number of observations.

Weighted least square (WLS):

The assumptions and the minimization term in WLS closely align with those in OLS. The only difference is that WLS accounts for heteroscedasticity by introducing a weighting term for individual OP observations, whose variance is assumed to be related to the variance of the residuals. The estimation equation in WLS is the same as that of OLS, but the objective to minimize is expressed as:

$$Minimize: \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 * w_i \; (Eq6)$$

$$w_i = \frac{1}{SD_i^{\;2}}$$

With w_i being the weight assigned to each observation, and SD_i is the OP analysis variance of each observation.

Positive least square (PLS):

The assumptions for PLS primarily include linearity, independence, and normality. PLS can be applied with weighting, if there is heteroscedasticity in the data. PLS extends OLS with the constraint that the regression coefficients must be non-negative. The estimation equation and the error term, PLS, are similar to OLS (without weighting) and WLS (applying weighting). To ensure the positivity of coefficients, a specific condition must be met:

$$\beta_i \geq 0, \forall i \; in \; PM \; sources \; (Eq7)$$

Ridge:

Shrinkage methods such as Ridge regression try to produce a more interpretable model or reduce error in the presence of collinearity by selecting a subset of the independent variables. Ridge regression is introduced by Hoerl & Kennard (1970), which incorporates a penalty term that shrinks the coefficients towards zero. The Ridge regression minimizes the residual sum of squares plus a penalty term proportional to the sum of squares of the coefficients (L2 regularization) as shown in Eq 8. Consequently, Ridge regression reduces the influence of a PM source that exhibits minimal impact on OP prediction without excluding it from the model.

$$Minimize: \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda * \sum_{j=1}^{p} \beta_j^2 \; (Eq8)$$

where λ is the parameter representing the amount of shrinkage, the larger λ, the greater the shrinkage. The hyperparameter tuning was implemented with different values of λ (5, 1, 0.5, 0.1, 0.01, 0.005, 0.001, 0.0005, 0.0001). The best λ for every site varied from 0.005 to 0.01 and in this study, 0.01 was selected. Ridge can be applied with weighting to account for heteroscedasticity.

Least Absolute Shrinkage and Selection Operator (Lasso):

Lasso (Tibshirani, 1996) is a shrinkage method that uses a penalty term proportional to the sum of the absolute regression coefficients (L1 regularization). This penalty term shrinks the coefficients of a source with a low impact on OP prediction to zero, effectively removing it from the model. This results in a sparse model that may be easier to interpret and may reduce error on out-of-sample data. However, Lasso is more sensitive to outliers than ridge regression and is less stable when data are collinear. Lasso can be applied with weighting to account for heteroscedasticity.

$$Minimize: \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda * \sum_{j=1}^{p} |\beta_j| \ (Eq9)$$

Similar to Ridge, λ is the parameter representing the amount of shrinkage. λ is selected as 0.01 in this study by running the hyperparameter tuning using the same values as for Ridge.

Generalized linear model (GLM)**:**

Generalized linear models, as introduced by McCullagh (1989), provide a framework for regression analysis that can contain non-normal error distributions and capture non-linear relationships between OP activities and PM sources. GLM allows for error variance that is a function of the predicted value, hence accounting for heteroskedasticity. Key assumptions underlying GLM include (1) independence, (2) the non-normal distribution of OP, and (3) the relationship between the PM sources and the transformed OP (logarithm in this study) is linear. The mathematical expression for GLM can be represented as follows:

$$log(\hat{y}) = \beta_0 + \sum_{0}^{p} \beta_i * x_i \ (Eq10)$$

where β_0 denotes the intercept, β_i represents the regression coefficient of source i, and x_i is the concentration of source i.

Random forest (RF):

RF, an ensemble learning method introduced by Breiman (2001), combines multiple decision trees to make predictions. In the reference implementation, each tree is grown on a bootstrap sample of the data and a random subset of the available features is evaluated at each node to choose the best split. The predictions of all trees are averaged to give the forest's final prediction. RF is customizable via hyperparameters such as the number of trees, the size of the bootstrap sample, and the number of features to evaluate at each node. The hyperparameters of RF in this study were chosen by tuning, as shown in section S1.1 Supplement.

RF does not assume a specific equation to express the relationship between OP activities and PM sources, with the result that intrinsic OP could not be computed in this regression model. Nevertheless, RF can estimate the relative importance of each PM source in OP prediction. This study estimated the permutation importance of each PM source as the mean increase in the mean squared error of predicted OP when the values of the PM source were permuted.

Multilayer perception (MLP):

MLP is an artificial neural network that consists of multiple layers of interconnected nodes or neurons organized in a feedforward structure (Akhtar et al., 2018; Bourlard & Wellekens, 1989; Chianese et al., 2018). These layers include an input layer (PM sources), one or several hidden layers, and an output layer (OP_{AA} or OP_{DTT} activities). In MLP, the neurons in the hidden layers are linked with the previous neurons by the connection weight, where every neuron is independent and has a different weight. The output of each neuron depends on its inputs and an activation function, which, if non-linear, allows the model to capture non-linear relationships. The implementation of MLP includes three steps: (1) forward pass to training model: the input is passed to the model, multiplied with an initial weight, add bias at every layer, then calculate output of the model. (2) error calculation: after applying step 1, the output of the model and the observed data are used to calculate the error. (3) backward pass: the error is propagated back through the network, and then the weights are adjusted to minimize overall error. These 3 steps are repeated until the error is minimized.

https://doi.org/10.5194/egusphere-2024-361
Preprint. Discussion started: 19 February 2024

The choice of hyperparameters to ensure the MLP model's robustness is processed by hyperparameter tuning and shown in section S1.2 of the supplement. Thanks to hyperparameter tuning, the two hidden layers and a logistic sigmoid activation function were selected in this study to capture the non-linear relationships between OP activities and PM sources.

All regression models were performed using the Python package statsmodels 0.14.0 (Seabold & Perktold, 2010) and scikit-learn 1.3.1 (Pedregosa et al., 2011).

2.6. Performance of the models

The performance metrics R-square (R^2), mean absolute error (MAE), and root mean square error (RMSE) were used to assess the goodness of fit of models as described by Kuhn & Johnson (2013). R^2 quantifies the model's ability to explain the variance in the data. R^2 equal to 1 indicates a perfect fit. RMSE represents the aggregation of the individual differences between predicted OP and measured OP, while MAE assesses the average magnitude of errors between them. Lower RMSE and MAE values indicate a better fit, with a perfectly fitting model yielding an RMSE or MAE of 0. Eq12, Eq13, and Eq14, respectively, define R^2, MAE, RMSE. These indicators are computed for the training and testing data of each sampling iteration and averaged across the 500 sampling iterations.

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sum of Squares}} = 1 - \frac{\sum_{i=0}^{m}(y_i - \hat{y_i})^2}{\sum_{i=0}^{m}(y_i - \hat{y_i})^2} \ (Eq12)$$

$$MAE = \frac{\sum_{i=0}^{m}(y_i - \hat{y_i})^2}{m} (Eq13)$$

$$RMSE = \frac{\sum_{i=0}^{m}(y_i - \hat{y_i})^2}{m} (Eq14)$$

3. Result and discussion

Assessments of collinearity and homoscedasticity are addressed in Section 3.1. Model performance, including key performance metrics and identification of the optimal model, is detailed in Section 3.2. Section 3.3 compares the intrinsic OP estimated by the different models. Section 3.4 compares intrinsic OP between the combined best-fit and reference models. Lastly, Section 3.5 proposes recommendations for selecting an appropriate model.

3.1. Dataset characteristics

The contributions of identified sources (µg m^{-3}) and the OP_v activities (nmol min^{-1} m^{-3}) in each site are presented in Figure 3, illustrating variations in annual average OP activities and PM source contributions by sites. Most sites, including traffic and industrial ones, show higher OP_{DTT} activities than OP_{AA}. Conversely, for the alpine valley sites, CHAM presents higher OP_{AA} than OP_{DTT}, while GRE-fr experiences similar levels of the 2 OPs. Additionally, the average OP activities in every site are not proportional to the average PM concentration. For instance, CHAM and NIC had lower PM concentrations but higher OP activities than other sites, while TAL showed high PM concentrations but relatively lower OP activities.

The variations observed in the levels of PM and OP across six sites can be attributed to distinctions in identified sources and their respective contributions. These disparities are contingent upon the unique typologies of each site, which are discussed in Weber et al., 2021. Further, we can observe a significant seasonality in the OP activities (Table S.1). Strong seasonality of OP in Alpine valley sites has been addressed in previous studies (Borlaza et al., 2021; Dominutti et al., 2023; Weber et al., 2018, 2021), with thermal inversions during winter increasing pollutants concentrations and OP activities compared to summer. Conversely, OP activities in cold and warm periods in other sites are not significantly different.

The PM sources and their repartition vary among sites (Figure 3) because of the difference in typology and local activities. For instance, in the industrial site (PdB), two specific sources are identified: shipping emissions (HFO) with an annual mean contribution of 1.39 µg m^{-3} and industrial sources at 0.86 µg m^{-3}. The urban background site TAL also appears to be influenced by industrial sources (2.34 µg m^{-3}), which might, however, be partly due to biases induced by the application of the harmonized receptor model protocol (Weber et al., 2019). Note that the application of a site-specific PMF procedure for this site leads to a much lower contribution of this source category but relatively similar contributions of other sources (Favez, 2017). GRE-fr, an urban background site in an alpine valley, presents significant long-range transport sources, with secondary sulfate contributing 3.90 µg m^{-3} followed by biomass burning at 2.21 µg m^{-3}. As expected, biomass burning is an abundant source in CHAM, accounting for 7.28 µg m^{-3} of the PM contribution, while the traffic sites RBX and NIC displayed high contributions of traffic sources (at 2.43 µg m^{-3} and 1.45 µg m^{-3} respectively).

The presence of multicollinearity and homoscedasticity were tested to assess the data characteristic of every site. The only site with evidence of collinearity was NIC, where the VIF of the traffic source was equal to 5.0. For all other sites, VIF values are below 5, indicating limited collinearity among sources. This is expected, as the PMF analysis is constrained to avoid collinearity between sources. VIF values for each site can be found in Table S.2.

The presence of heteroscedasticity is commonly found when the dependent variable (or OP in this study) exhibits a large difference between the minimum and maximum values or when the error variance varies proportionally with an independent variable (PM sources). The heteroscedasticity was assessed by applying the Goldfeld–Quandt test. Table 2 presents the p-values of the Goldfeld–Quandt test, indicating homoscedasticity of OP prediction when $p > 0.05$. This test reveals that heteroscedasticity was detected in CHAM, GRE-fr, NIC for OP_{AA} and in CHAM and TAL for OP_{DTT} (Table 2). We observed a large difference between the cold and warm periods for both OP_{AA} and OP_{DTT} in CHAM, similar to what was seen for OP_{AA} in GRE-fr (Table S1), which can be the reason for the presence of heteroscedasticity. For NIC and TAL, there is an insignificant difference between the cold and warm periods, which indicates the presence of heteroscedasticity may be because of the relationship between the PM sources and error variance. When heteroscedasticity is detected, unweighted regression for OP prediction according to sources may not accurately reflect the uncertainty of each source's intrinsic OP. The scatterplots representing the relationship between the regression analysis residuals and the fitted values (for observed OP) are available in Figures S.1 and S.2, Supplement.

Table 2. The p-value of the Goldfeld–Quandt heteroscedasticity test

	PdB	**TAL**	**GRE-fr**	**CHAM**	**RBX**	**NIC**
AA	0.15	0.78	<< 0.001	<< 0.001	0.44	0.002
DTT	0.59	<< 0.001	0.189	<< 0.001	0.56	0.91

https://doi.org/10.5194/egusphere-2024-361
Preprint. Discussion started: 19 February 2024

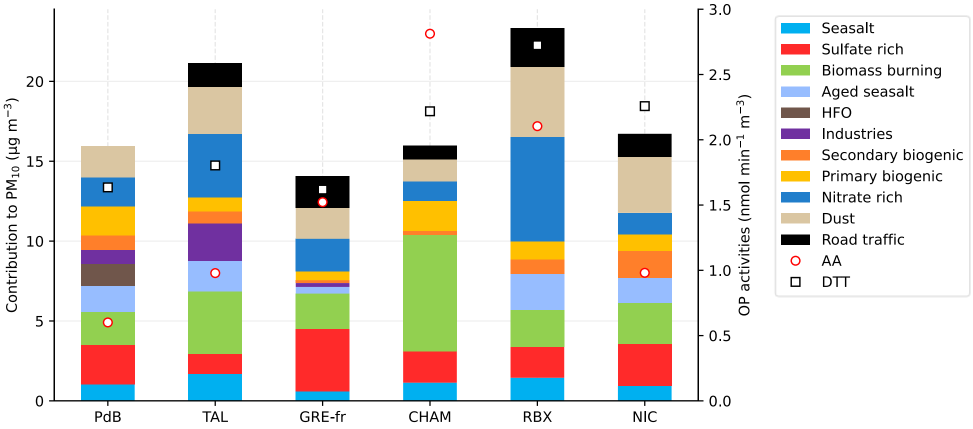


Figure 3. The contribution of sources to PM$_{10}$ and the OP activities in 6 sites. The left y-axis and bar show the contribution of PM sources in µg m^{-3}. The right y-axis, circles and squares showed the mean OP$_v$ activities in nmol min^{-1} m^{-3}, with red circle for OP$_{AA}$ and black square for OP$_{DTT}$.

3.2. The performances of regression models

The 11 regression models, with or without weighing for some of them, were tested by comparing their performance metrics between the measured and reconstructed OPs. For each run (n = 500 iterations), the R^2, RMSE, and MAE were computed for the testing and training dataset, resulting in 500 values for each performance metric. Figure 4 presents the mean R² values of the training data sets, the mean and the standard deviation of the testing datasets of the OP$_{AA}$ models across the 500 sampling iterations, and Figure 5 presents the mean RMSE and MAE. The same result pattern was found for OP$_{DTT}$, as presented in the tables S.3, S.4, S.5, Supplement. The WLS, wPLS, wRidge, and wLasso models incorporated weighting, while the OLS, PLS, Ridge, Lasso, GLM, RF, and MLP models were unweighted.

https://doi.org/10.5194/egusphere-2024-361
Preprint. Discussion started: 19 February 2024

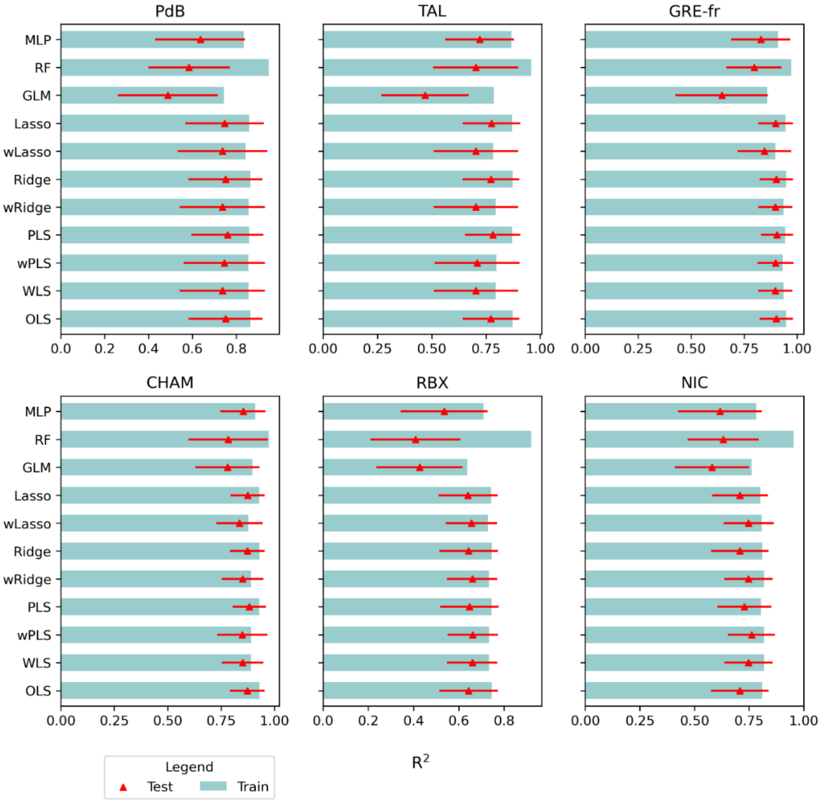


Figure 4. The R^2 of 11 OP$_{AA}$ models in 6 sites. The mean R^2 of training data is shown in a blue bar, the mean R^2 of testing data is shown by a red triangle, and the red bar is the standard deviation of the R^2 of the testing data. The y-axis represents the models, and the x-axis denotes the R^2 values. 360

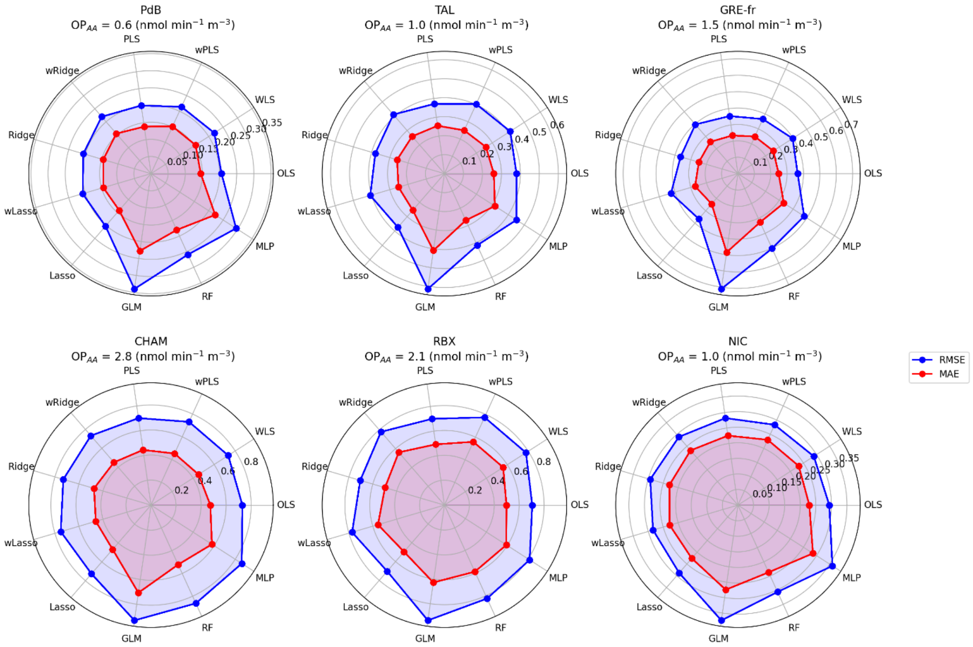


Figure 5. The MAE and RMSE of 11 OP_{AA} models in every site for the testing data. Blue and red lines present the RMSE and the MAE, respectively. The values in the figure are the mean of RMSE and MAE of 500 iterations.

OP predictions across all sites are statistically validated, with testing R^2 values observed in RBX, NIC, PdB, TAL, CHAM, and GRE-fr being 0.66, 0.76, 0.76, 0.78, 0.87, 0.90, respectively. The lowest mean test set RMSE values are 0.70, 0.28, 0.21, 0.37, 0.70, 0.31 nmol min^{-1} m^{-3}, respectively, for the same sites. The lowest mean test set MAE values are 0.49, 0.23, 0.14, 0.25, 0.45, and 0.21 nmol min^{-1} m^{-3}, respectively. Notably, the GLM model exhibits for all sites the lowest R^2 values and the highest RMSE (Table S.3, S.4, S.5, Supplement). These results strongly suggest that the relationship between OP_{AA} and PM sources is not log-linear.

Differences in MAE, RMSE, and R^2 between the training and testing database for RF and MLP are significant across the sites. Notably, RF displays a large difference in R^2, with a gap of up to 0.6 in RBX (R^2 training: 0.92, R^2 testing: 0.27). Similar gaps were found in RMSE and MAE. RF consistently performed best on the training set, characterized by the highest R^2 and the lowest MAE and RMSE values, but had lower set test R^2 values than the other models (except GLM). Conversely, MLP exhibited training R^2 values comparable to other models but lower test R^2. These findings suggest overfitting: the flexible algorithms identify relationships in the training data that do not generalize to the testing data. This observation may be attributed to the limitations of data coverage, possibly failing to fully represent the underlying relationships, leading to poor performance in testing datasets (Benkendorf & Hawkins, 2020; Hawkins, 2004; Hernandez et al., 2006; Matsuki et al., 2016; Raudys & Jain, 1991; Stockwell & Peterson, 2002; Wisz et al., 2008). Pearce and Ferrier (2000) recommended that the minimum number of samples for robust performance should be over 250 for GLM model, while (Raudys & Jain, 1991) showed that the minimum number of sample are based on the complexity of the model and the number of predictors. Additionally, Harrell (2016) suggested that the number of predictors (PM sources) should be below the number of samples divided by 15, a threshold not reached in this analysis. For example, in NIC, the minimum number of samples should be 135 for the training set (9 PM sources x 15), while in total, we have only 107 samples. Therefore, we can also recommend that, for optimal performance of RF, and MLP, the number of samples and PM sources should satisfy these thresholds.

The WLS, OLS, wPLs, wRidge, and wLasso models show more robust performances with fewer differences between the training and testing data. At most sites, there is very little difference between the R^2, RMSE, and MAE of OLS and Ridge, with or without weighting, and often PLS and Lasso as well. This consistency is observed even in the collinearity case of NIC, where VIF = 5. The difference between these models is a maximum of 0.06 in R^2, 0.01 in MAE and 0.1 in RMSE, indicating that these models work well for OP prediction. Nevertheless, it is worth noting that every model exhibits different assumptions that have to be respected. The assumption violations may lead to unreliable regression coefficients (intrinsic OP) even though the prediction is good (Cohen et al., 2013; Williams et al., 2013).

The best model for each site was selected based on both data characteristics (collinearity and heteroscedasticity) and testing data performance. For sites with collinearity, the Ridge, Lasso were considered most appropriate. For sites with heteroscedasticity, models with weights were considered the most appropriate. For sites with neither collinearity nor heteroskedasticity, OLS and PLS were considered most appropriate. Tables 3 and 4 present the best OP_{AA} and OP_{DTT} prediction models for each site. It follows that the best model is not necessarily the same one for both series of OP for a given site. As a rule, the model that exhibits the best performance metrics (the best model by error in Table 3 for OP_{AA} and Table 4 for OP_{DTT}) is suited to the best model chosen by data characteristics; therefore, choosing a model according to data characteristics help to more reliable in OP predictions.

Table 3. Criteria to select the best model for OP_{AA}

	PdB	TAL	GRE-fr	CHAM	RBX	NIC
Collinearity	No	No	No	No	No	Yes
Heteroscedasticity	No	No	Yes	Yes	No	Yes
Best model by characteristic	OLS/ PLS	OLS/ PLS	WLS/ wPLS	WLS/ wPLS	OLS/ PLS	wRidge/ wLasso
Best by error	PLS	PLS	wPLS	wPLS	OLS	wRidge

Table 4. Criteria to select the best model for OP_{DTT}

	PdB	TAL	GRE-fr	CHAM	RBX	NIC
Collinearity	No	No	No	No	No	Yes
Heteroscedasticity	No	Yes	No	Yes	No	No
Best model by characteristic	OLS/ PLS	WLS/ wPLS	OLS/ PLS	WLS/ wPLS	OLS/ PLS	Ridge/ Lasso
Best by error	OLS	wPLS	PLS	wPLS	PLS	Ridge

3.3. Effect of the choice of a model on intrinsic OP

It is particularly important to try to define the best way of calculating the more accurate PM sources intrinsic OP and the contribution of sources to OP, since these values are fundamental inputs in all the works of large-scale modelling of OP with chemical transport models (CTM) (Daellenbach et al., 2020; Vida et al., 2024). Figures 6 and 7 show the variations of intrinsic OP for all the models, focusing on the results of NIC as an example. The evaluation of the 5 other sites is presented in Fig S.3 to Fig S.7 for OP_{AA} and Fig S.8 to S.12 for OP_{DTT}. The differences in equations, error term minimizations, and assumptions can explain the differences in intrinsic OP per µg of source among the eight regression models. While the R², RMSE, and MAE values are similar among models (except for GLM, RF, and MLP), the intrinsic OP values significantly differ between the models with and without weighting and between the linear and non-linear regression models. The average intrinsic OP of 500 iterations is discussed in this section since these values are usually used to calculate the contribution of the PM source to OP in prior studies (Borlaza et al., 2021; Dominutti et al., 2023; Weber et al., 2018). The mean and standard deviation of intrinsic OP_{AA} and OP_{DTT} for the 6 sites are shown in Table S.6 and S.7, respectively.

Intrinsic OP_{AA} of PM sources at NIC is the same between WLS and wRidge and between the OLS and Ridge, revealing that the moderate collinearity of the road traffic source did not affect the estimated intrinsic OP_{AA}. PLS sets the intrinsic OP_{AA} of some sources to zero, therefore producing slightly different results. Lasso regression sets the intrinsic OP_{AA} of some sources to zero and shrinks the estimates for all other sources toward zero. GLM produces intrinsic OP_{AA} values that represent a multiplicative change on the log scale, so they are not directly comparable to the other models. However, the direction and importance of the sources are similar to the other models. Whatever the model, road traffic appears as the source with the highest intrinsic OP_{AA}, followed by biomass burning, aged salt, salt and sulfate-rich sources, in NIC. Traffic and biomass burning sources have been similarly recognized as significant contributors to OP_{AA} in prior studies (Borlaza et al., 2021; Dominutti et al., 2023; Stevanović et al., 2023). The intrinsic OP of the dominant sources is stable, indicating that all these models could give the same information about the intrinsic OP of the main sources. Conversely, the differences are larger between models for the sources with small to very small intrinsic OP (MSA rich, primary biogenic, nitrate-rich, dust), whose intrinsic OP varies from positive to negative among models.

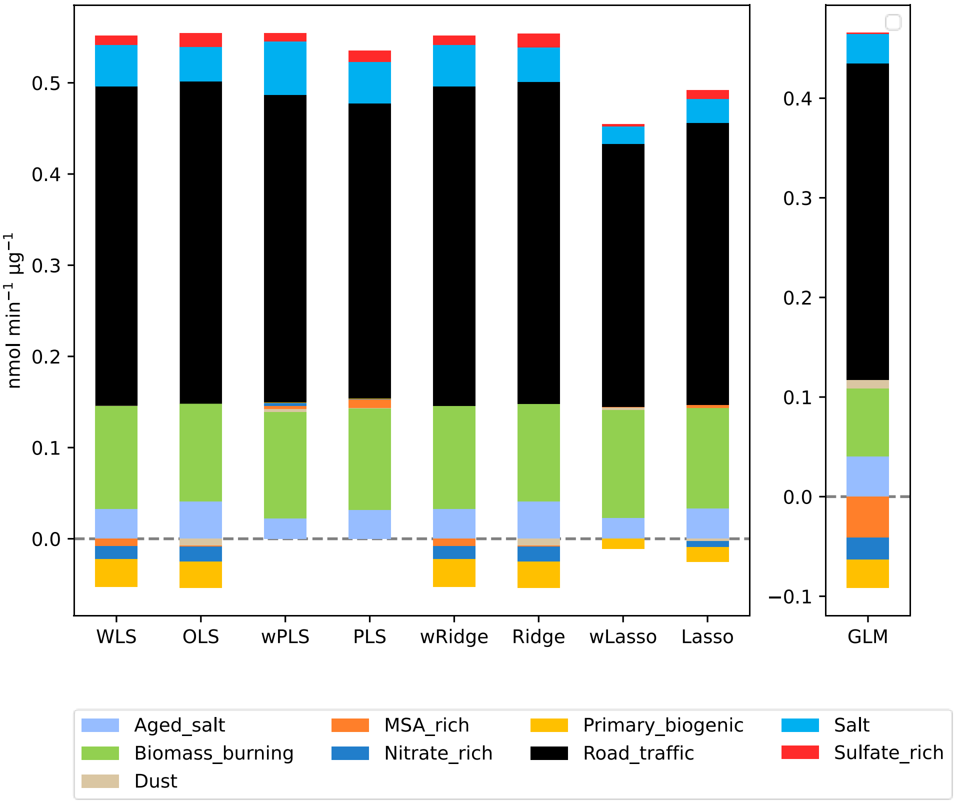


Figure 6. Intrinsic OP_{AA} values of the different PM sources at Nice were obtained with the different models.

https://doi.org/10.5194/egusphere-2024-361
Preprint. Discussion started: 19 February 2024

The OP_{DTT} intrinsic values in NIC (Figure 7) display minimal variation among the WLS, wPLS. This consistency is linked to the absence of negative intrinsic values. On the other hand, even though there is the presence of moderate collinearity, wRidge still has the same result as WLS and wPLS. In line with the OP_{AA} results, the wLasso and GLM models exhibit distinct responses compared to the other models. The intrinsic OP_{DTT} of all sources varies depending on the presence or absence of weighting. While the WLS models tend to amplify the influence of some sources (aged sea salt, primary biogenic, sea salt, and sulfate-rich), the OLS reduces the intrinsic OP_{DTT} of these sources. Conversely, MSA-rich, nitrate, and road traffic sources undergo less influence in WLS but higher in OLS. Different from OP_{AA}, OP_{DTT} prediction shows more variation among models, highlighting the effect of choosing a model on evaluating the intrinsic OP_{DTT} of PM sources.

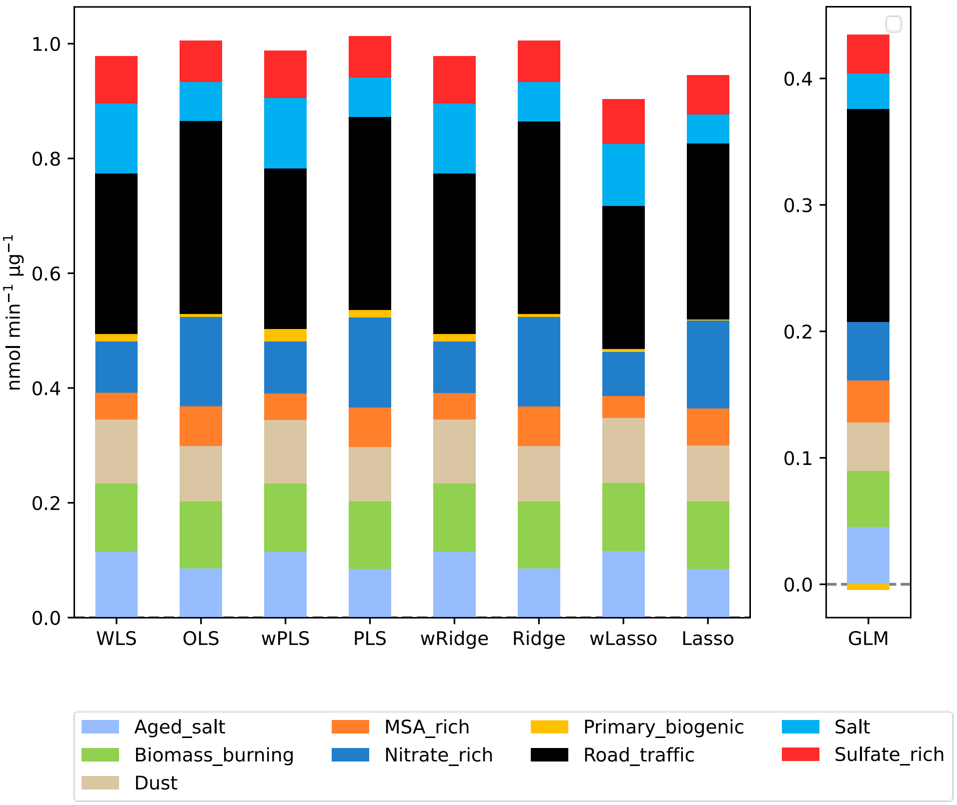


Figure 7. The variations of the intrinsic OP_{DTT} of the different PM sources at Nice were obtained with the different models.

The comparison of intrinsic OP among regression models in NIC demonstrated that OP_{DTT} and OP_{AA} intrinsic values exhibit variation across different models with and without weighting, illustrating that the choice of the model significantly influences the values obtained for intrinsic OP of PM sources (A similar pattern is observed for all other sites and shown in Fig S.3 to Fig S.7 for OP_{AA} and Fig S.8 to S.12 for OP_{DTT}). Because of the difference in OP intrinsic across models, a comparison between the best-performing and most commonly used models (OLS) is presented in the following section to elucidate the advantage of choosing a model based on data characteristics (section 3.4).

3.4. Comparisons between the best site-specific model and OLS

In this section, the intrinsic OP of the best model is selected for each site as discussed in Section 3.2, and the intrinsic values of each source are compared to the ones returned by the OLS model. The OLS model is used as a representative of usual practices that do not consider the database characteristics. Each PM source's average intrinsic OP value is calculated from all the 500 bootstrapping iterations for all sites where that particular source is identified. Intrinsic OP values obtained in this way from the best model encompassing all six sites are called **intrinsic OP of the best model,** and the intrinsic OP values derived from the OLS from all six sites are called **intrinsic OP of the reference model.**

A meaningful comparison of the two series of intrinsic values requires two conditions. First, intrinsic OP values should be consistent across all sites. While recognizing that intrinsic OP values depend on diverse factors, we assumed the sites share fairly uniform PM chemical source profiles in France. This is demonstrated by evaluating the Pearson distance and standardized identity distance similarity indicators of the source chemical profiles (Belis et al., 2015; Weber et al., 2019), and Figure S.13 indicates consistent profiles of sources for the 6 sites. Consequently, we could expect to observe minimal divergence in intrinsic OP values among these sites. Second, we postulate that negative intrinsic OP values are possible since previous studies have reported that total PM intrinsic OP can be modulated due to the synergetic/antagonistic effects involving, for example, soluble copper, quinones, and bacteria (Borlaza et al., 2021; Pietrogrande et al., 2022; Samake et al., 2017). These last studies showed that the impact of synergistic and antagonistic effects cannot exceed 60% of the intrinsic OP value when assessed independently for each chemical. Consequently, we consider here that the intrinsic OP value of an individual site for a given source could be negative only within a range of at most 60% of the mean combined intrinsic OP value of this source across all sites. Negative intrinsic OP exceeding this criterion may result from the mathematical construction of the model. The comparison of intrinsic OP_{AA} of the best and reference model is presented in 3.4.1 and that of OP_{DTT} is shown in 3.4.2.

3.4.1. OP_{AA} activities

The results of the comparison of OP_{AA} intrinsic values (Figure 8 and Table S.8) show that the anthropogenic sources get the highest intrinsic OP values in both the best and reference models. Among these sources, road traffic appears as the most prominent potent fraction, followed by biomass burning, HFO, and industrial. These results are aligned with prior research (Calas et al., 2019; Daellenbach et al., 2020; Dominutti et al., 2023; Fadel et al., 2023; Fang et al., 2016; in ’t Veld et al., 2023a; Weber et al., 2018; Zhang et al., 2020) which has highlighted the sensitivity of OP_{AA} to concentrations of metals, black carbon, and organic carbon. The differences between the best and reference models were insignificant for these sources, demonstrating that **the best and reference models consistently captured similar patterns for the most critical sources of OP activities**.

However, the interquartile ranges (IQR) of the intrinsic OP values are consistently narrower for the best models across all sources, accounting for less divergence in intrinsic OP values across sites. Moreover, the median intrinsic OP values obtained from the best model closely approximated the mean values, indicating the absence of extreme intrinsic OP values. For instance, in the case of road traffic, the mean and median values were 0.24 and 0.23 nmol min^{-1} μg^{-1}, respectively. Conversely, the reference model exhibited a large difference between the mean and median values, implying lower consistency across sites and sampling iterations. The same result was observed in biomass burning source, in which the median and mean intrinsic OP in the best model had fewer discrepancies. Further, the biomass burning intrinsic OP in GRE-fr of the best model is more consistent with those in other sites (best: 0.30 nmol $min^{-1}\mu g^{-1}$, reference: 0.35 nmol $min^{-1}\mu g^{-1}$).

When considering sources with low intrinsic OP, the variability can be larger between the two methods. As an example, for the sulfate-rich sources, the median intrinsic OP values were positive (0.002 nmol min^{-1} μg^{-1}), while the mean intrinsic OP values were negative (-0.008 nmol $min^{-1}\mu g^{-1}$). The mean intrinsic OP in the best model

https://doi.org/10.5194/egusphere-2024-361
Preprint. Discussion started: 19 February 2024

500 exhibited fewer negative values in individual sites than in the reference model (for aged salt, salt, primary biogenic, MSA rich, sulfate-rich and nitrate-rich), highlighting the advantage of considering the data in model selection. For example, the mean intrinsic OP values of the primary biogenic source revealed a negative intrinsic OP in NIC (-0.03). This negative value represented a 100% reduction compared to the mean intrinsic OP of all sites. In the OLS model, the negative intrinsic OP observed in NIC and some extreme values in GRE-fr, CHAM, NIC (where 505 heteroscedasticity was presented) underscores that the model assumptions on data characteristics proving false could impact the accuracy of OP prediction.

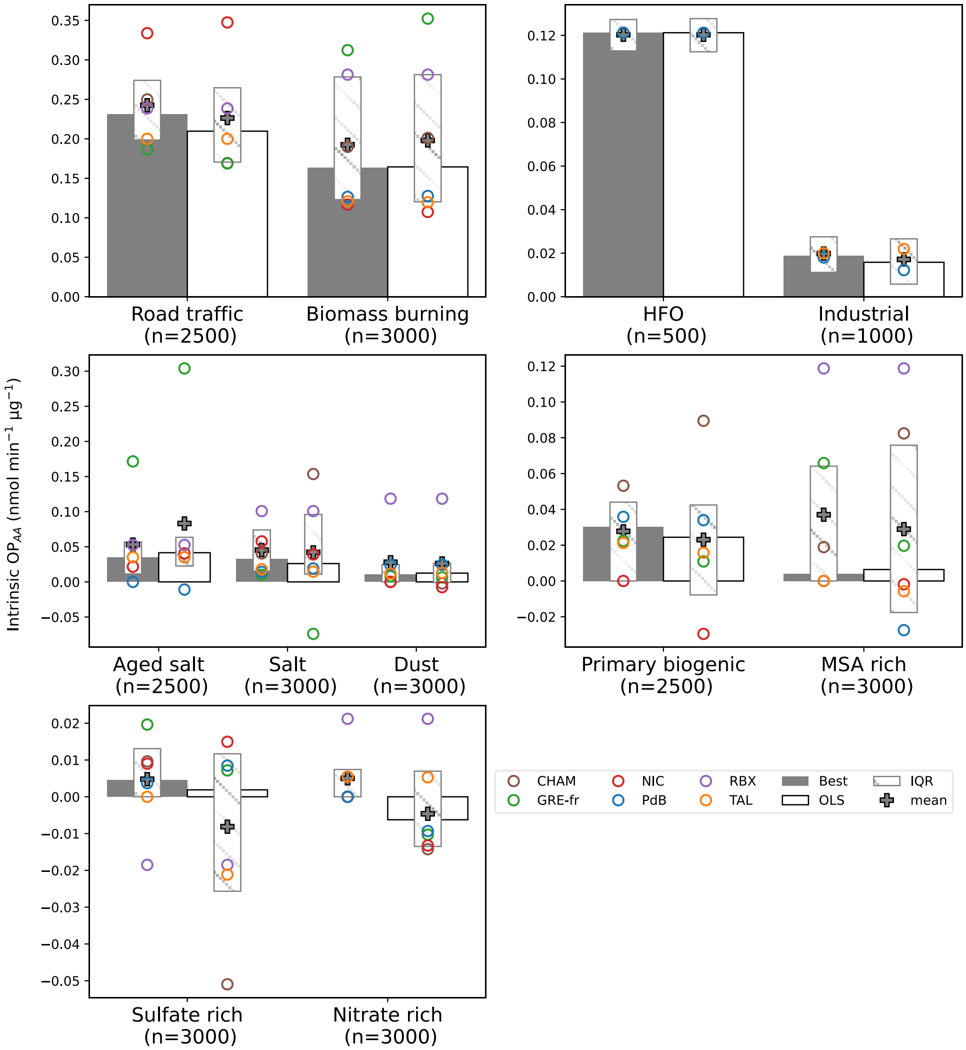


Figure 8. Intrinsic OP$_{AA}$ estimated by the best and the reference methods in the 6 sites. The y-axis represents the intrinsic OP values in nmol min^{-1} µg^{-1}, the x-axis represents the sources. The grey bars are the median 510 intrinsic OP values of the best models in the 6 sites (n = 500 bootstrapping * number of sites where the given source is detected) for each source. The white bars are the same median intrinsic OP values for the reference (OLS) model. The grey plus symbol represents the mean of OP intrinsic values. The hatched bars are the interquartile ranges of the intrinsic OP values. The dots represent the mean intrinsic OP of all sites,

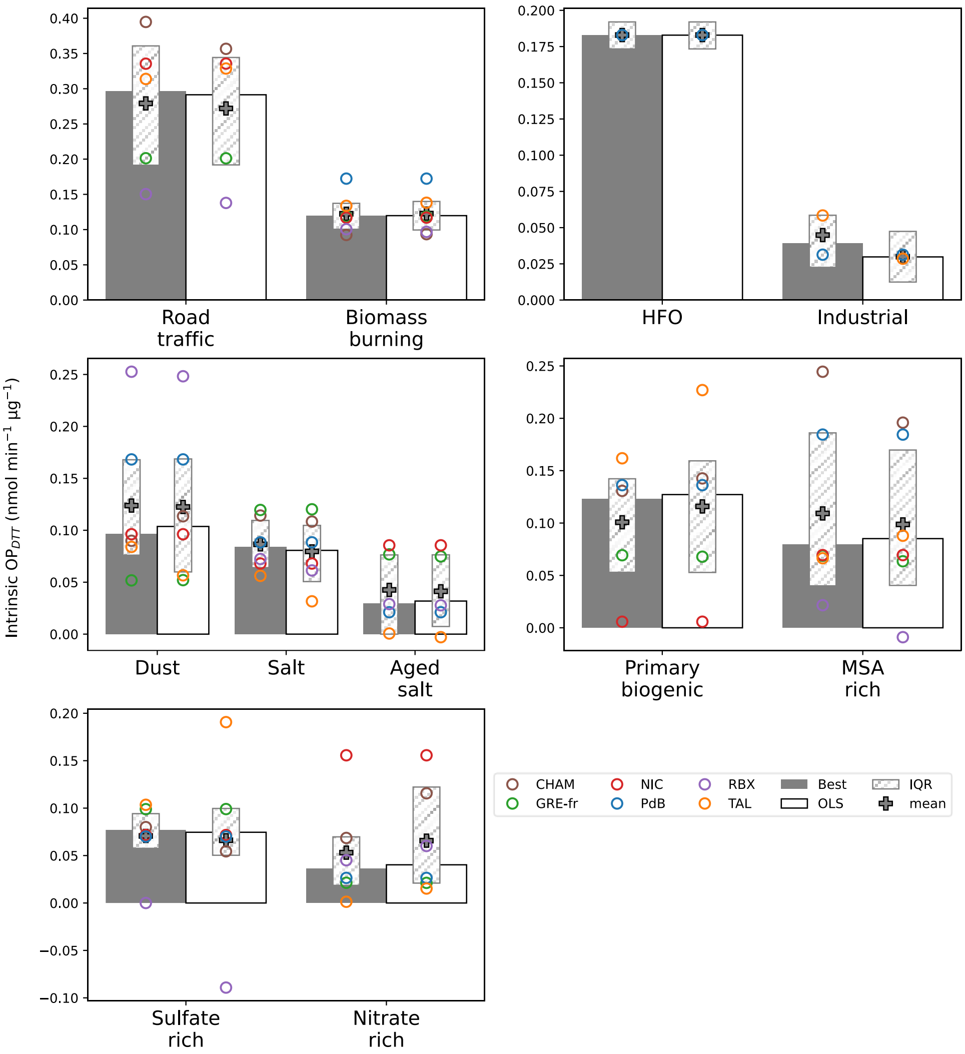
including grey – Chamonix, green – Grenoble, red – Nice, blue – Port-de-Bouc, purple – Roubaix, and orange-Talence.

The detailed comparison of intrinsic OP_{AA} between the best and reference models is categorized into four groups and discussed in detail in section S9. These groups include (1) anthropogenic sources without nitrate and sulfate (road traffic, biomass burning, HFO, industrial), (2) natural inorganic sources (aged sea salt, sea salt, dust), (3) biogenic sources (primary biogenic, MSA rich), and (4) nitrate and sulfate-rich sources.

3.4.2. OP_{DTT} activities

Similar to OP_{AA}, for OP_{DTT} the IQR of the best model is narrower for most of the sources than the IQR of the reference model (OLS). Except for the road traffic, industrial, and MSA-rich, the IQR is slightly higher in the best model (Figure 9 and Table S.9). In the two models, the mean intrinsic OP is essentially unchanged, where the traffic is the most critical source (0.27±0.10), followed by HFO (0.18±0.01), biomass burning (0.12±0.03), dust (0.12±0.07), primary biogenic (best: 0.10±0.06, reference: 0.12±0.08) and MSA rich (best: 0.11±0.09, reference: 0.09±0.09). The remaining sources, such as sea salt, sulfate-rich, industrial, and nitrate-rich, show a negligible contribution to OP_{DTT} with an intrinsic OP_{DTT} from 0.02 to 0.08. The minimum difference between the two models again confirms the conclusion in the OP_{AA} comparison, demonstrating **the similar pattern of the best and the reference model in the most crucial sources of OP**. For both best and reference, OP_{DTT} activities showed sensitivity to more sources than OP_{AA}, as discussed in many works (Borlaza et al., 2021; Calas et al., 2019; Dominutti et al., 2023; Fadel et al., 2023). The traffic, HFO and biomass burning sources highlighted in (Fadel et al., 2023; Veld et al., 2023b; Serafeim et al., 2023; Y. Wang et al., 2020) are the most contributing to OP_{DTT} activities. The primary biogenic highly contributes to OP_{DTT} in both models, likely reflecting the sensitivity of OP_{DTT} to organic compounds, as mentioned in (Dominutti et al., 2023; Li et al., 2023; Weber et al., 2021). The intrinsic OP of dust and MSA-rich have been shown to vary in the literature, indicating the different effects of the compositions to generate DTT ROS.

While the best and reference models give the same mean intrinsic OP_{DTT} of all sites, the mean OP_{DTT} at each individual site can vary substantially between the two models. For the sulfate-rich source, the reference model showed a negative intrinsic OP_{DTT} in RBX (-0.09 nmol min^{-1} μg^{-1}), while the best model showed an intrinsic of 0. On the other hand, the reference model presents that the intrinsic OP in TAL is 0.20 nmol min^{-1} μg^{-1}, far from the mean of all sites (0.07 nmol min^{-1} μg^{-1}). The best model, conversely, shows a more consistent intrinsic OP in TAL compared to the other sites. A similar result was also found in primary biogenic sources, where the reference model overestimates the intrinsic OP of this source in TAL compared to the other sites. The reason is that heteroscedasticity was detected in TAL, which does not satisfy the assumption of OLS.



545

Figure 9. Intrinsic OP_{DTT} was estimated by the best and the reference methods in the 6 sites. The y-axis represents the intrinsic OP values in nmol min^{-1} µg^{-1}, the x-axis represents the sources. The grey bars are the median intrinsic OP values of the best models in the 6 sites (n = 500 bootstrapping * number of sites where the given source is detected) for each source. The white bars are the same median intrinsic OP values for the reference (OLS) model. The grey plus symbol represents the mean of intrinsic OP values. The hatched bars are the interquartile ranges of the Intrinsic OP values. The dots represent the mean intrinsic OP of all sites, including grey – Chamonix, green – Grenoble, red – Nice, blue – Port-de-Bouc, purple – Roubaix, and orange-Talence.

The comparison of intrinsic OP between the best models and the reference model highlights the importance of considering the database characteristics when selecting a model for OP SA. For all the datasets studied here, using the best model for each site delivered more robust results with reduced uncertainty, reduced differences in intrinsic

OP across sites, and provided a more geochemically meaningful intrinsic OP. The recommendation for selecting a model based on the characteristics of the database is presented in section 3.5.

3.5. Guidelines for the selection of regression model for OP SA.

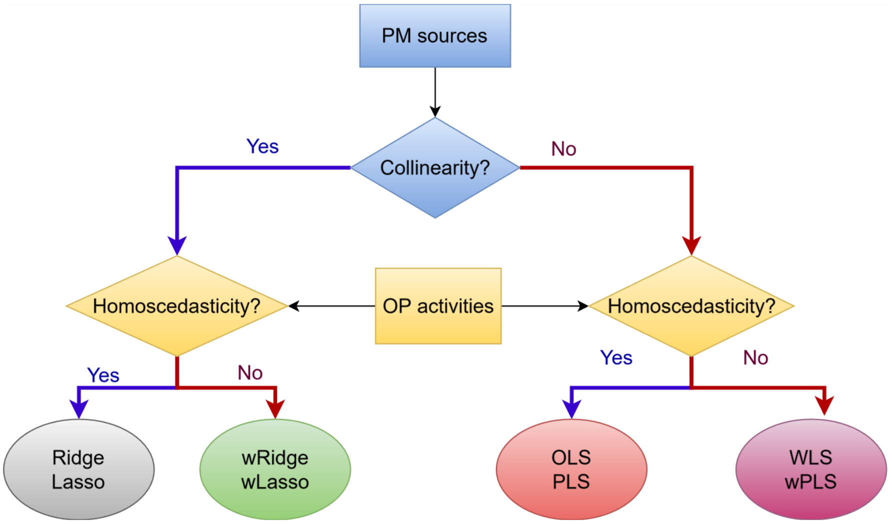


Figure 10. Workflow in model selection considering the characteristics of data

Our results have highlighted the benefits of choosing a model that matches the characteristics of the data to improve the robustness of OP SA method. For this reason, this section develops a workflow to help make model selection decisions. Before selecting a regression for OP SA, the first question is whether the PM sources are collinear and the second is whether the residual variance of the regression between OP and PM mass is constant. These two questions represent the characteristics of PM sources and OP activities, which vary according to the study site.

For data exhibiting collinearity between sources and generating a residual variance that varies according to the value of the PM sources, weighted regularisation regression can help to reduce collinearity and to match the model assumption about the residual. On the other hand, the unweighted Ridge and Lasso are introduced for data showing collinearity and homoscedasticity. Additionally, data with no collinearity are suitable for OLS and unweighted PLS in the case of homoscedasticity, while WLS, weighted PLS are used for data with heteroscedasticity.

If the number of predictors (PM sources) is below the number of samples divided by 15, RF and MLP can also be employed to capture possible non-linear relationships between the OP and PM sources. However, cross-validation must be used to ensure that there is no over-fitting. In addition, these models do not estimate intrinsic OP (nmol min^{-1}µg^{-1}) but only the importance of each PM source to the OP prediction. This is a large drawback since the intrinsic OP of sources is a must for the modelling effort of OP with CTM. However, RF and MLP could be useful for OP prediction in the case of larger datasets generated by online instruments.

For each data characteristic there is more than one model that suits. Out-of-sample performance metrics should be employed to identify the most accurate of these models.

Limitations and perspectives of the study

- This study compares eight regression models but is not exhaustive; further research could add more regression techniques to evaluate result variations across models.

- PMF coupled with a regression model remains a popular approach for OP SA. Notably, the uncertainties in PMF are typically addressed in chemical profiles, but not in contributions. Incorporating uncertainty from variations in contribution into models could enhance their robustness compared to relying only on absolute PMF results.
- Observations ranged between 100 and 200 samples at each site, which may be insufficient to obtain fair performance of GLM, decision trees and neural network models. Such a number of samples is sufficient to address SA through PMF model for offline analyses. Therefore, such study outlines well the limitations of GLM, RF, MLP for such types of datasets. Future investigations should be performed in an extended dataset, such as long-term or real-time measurement data, to investigate the performance of such machine learning algorithms.

4. Conclusion

The results of the OP SA marked an important milestone as they were revealed for the first time through the use of eight regression models, including OLS, WLS, PLS, GLM, Ridge, Lasso, RF and MLP. This in-depth analysis was carried out on a complete set of data collected from six sites with different characteristics. The approach of selecting a suitable model for each site based on specific data characteristics resulted in a more consistent intrinsic OP across sites, in stark contrast to the variation observed when using the basic OLS model. The revelations of the study have provided concrete recommendations for the judicious selection of an appropriate regression model based on the unique characteristics of the dataset. These guidelines should help to improve the accuracy of OP assessments and contribute to the refinement of air quality assessment methods. In addition, the implications of this research extend to the implementation of OP monitoring as a new measure of air quality, particularly on European supersites. As this initiative aligns with the ongoing revision process of the European Directive 2008/50/CE, the study's findings assume a pivotal role in shaping the methodologies underpinning air quality assessments at a broader regulatory level.

Code availability

The software code could be made available by contacting the corresponding author upon request.

Data availability

The datasets could be made available upon request by contacting the corresponding author.

Author contributions

VDNT performed the data analysis for the OP source apportionment setup. GU, JLJ mentoring, supervision, and validation of the methodology and results. IH, PD, and VDNT worked on the result visualization. OF, JLJ, and GU acquired fundings for the original PM sampling and analysis. VDNT wrote the original draft. All authors reviewed and edited the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

The authors would like to express their sincere gratitude to many people of the Air-O-Sol analytical platform at IGE (including S. Darfeuil, R. Elazzouzi, and T Madhbi), to R. Aujay (Ineris) for sample management at TAL and RBX, to L. Alleman (IMT Nord-Europe) and N. Bonnaire (LSCE) for part of the chemical analyses for some sites, and to all the personnel within the AASQA in charge of the sites for their contribution in conducting the dedicated sample collection. The authors would like to thank S. Weber for running the PMF model in his previous professional life.

Financial support

The PhD grant of VDNT was funded by grant PR-PRE-2021, UGA-UGA 2022-16 FUGA-Fondation Air Liquide, and ANR ABS (ANR-21-CE01-0021-01). Analytical work on OP was funded through ANR GET OP STAND (ANR-19-CE34-0002), MOBILAIR and ACME IDEX projects at UGA (ANR-15-IDEX-02). The sampling and chemical analyses performed at TAL, GRE, RBX, PdB and NIC sites have been partly funded by the French Ministry of Environment in the frame of the CARA program. The present work was also supported by European Union's Horizon 2020 research and innovation program under grant agreement 101036245 (RI-URBANS) for the Post-doc salary of Pamela Dominutti.

Reference

Akhtar, A., Islamia, J. M., Masood, S., Islamia, J. M., Masood, A., & Islamia, J. M. (2018). *Prediction and Analysis of Pollution Levels in Delhi Using Multilayer Perceptron*. *June*. https://doi.org/10.1007/978-981-10-3223-3

Akhtar, McWhinney, R. D., Rastogi, N., Abbatt, J. P. D., Evans, G. J., & Scott, J. A. (2010). Cytotoxic and proinflammatory effects of ambient and source-related particulate matter (PM) in relation to the production of reactive oxygen species (ROS) and cytokine adsorption by particles. *Inhalation Toxicology*, *22*(SUPPL. 2), 37–47. https://doi.org/10.3109/08958378.2010.518377

Alleman, L. Y., Lamaison, L., Perdrix, E., Robache, A., & Galloo, J. C. (2010). PM10 metal concentrations and source identification using positive matrix factorization and wind sectoring in a French industrial zone. *Atmospheric Research*, *96*(4), 612–625. https://doi.org/10.1016/j.atmosres.2010.02.008

Ayres, J. G., Borm, P., Cassee, F. R., Castranova, V., Donaldson, K., Ghio, A., Harrison, R. M., Hider, R., Kelly, F., Kooter, I. M., Marano, F., Maynard, R. L., Mudway, I., Nel, A., Sioutas, C., Smith, S., Baeza-Squiban, A., Cho, A., Duggan, S., & Froines, J. (2008). Evaluating the toxicity of airborne particulate matter and nanoparticles by measuring oxidative stress potential - A workshop report and consensus statement. *Inhalation Toxicology*, *20*(1), 75–99. https://doi.org/10.1080/08958370701665517

Bates, J. T., Weber, R. J., Abrams, J., Verma, V., Fang, T., Klein, M., Strickland, M. J., Sarnat, S. E., Chang, H. H., Mulholland, J. A., Tolbert, P. E., & Russell, A. G. (2015). Reactive Oxygen Species Generation Linked to Sources of Atmospheric Particulate Matter and Cardiorespiratory Effects. *Environmental Science and Technology*, *49*(22), 13605–13612. https://doi.org/10.1021/acs.est.5b02967

Bates, J. T., Weber, R. J., Verma, V., Fang, T., Ivey, C., Liu, C., Sarnat, S. E., Chang, H. H., Mulholland, J. A., & Russell, A. (2018). Source impact modeling of spatiotemporal trends in PM2.5 oxidative potential across the eastern United States. *Atmospheric Environment*, *193*(August), 158–167. https://doi.org/10.1016/j.atmosenv.2018.08.055

Beelen, R., Stafoggia, M., Raaschou-Nielsen, O., Andersen, Z. J., Xun, W. W., Katsouyanni, K., Dimakopoulou, K., Brunekreef, B., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Houthuijs, D., Nieuwenhuijsen, M., Oudin, A., Forsberg, B., Olsson, D., Salomaa, V., Lanki, T., … Hoek, G. (2014). Long-term exposure to air pollution and cardiovascular mortality: An analysis of 22 European cohorts. *Epidemiology*, *25*(3), 368–378. https://doi.org/10.1097/EDE.0000000000000076

Belis, C. A., Karagulian, F., Amato, F., Almeida, M., Artaxo, P., Beddows, D. C. S., Bernardoni, V., Bove, M. C., Carbone, S., Cesari, D., Contini, D., Cuccia, E., Diapouli, E., Eleftheriadis, K., Favez, O., El Haddad, I., Harrison, R. M., Hellebust, S., Hovorka, J., … Hopke, P. K. (2015). A new methodology to assess the performance and uncertainty of source apportionment models II: The results of two European intercomparison exercises. *Atmospheric Environment*, *123*, 240–250. https://doi.org/10.1016/j.atmosenv.2015.10.068

Belis, C. A., Karagulian, F., Larsen, B. R., & Hopke, P. K. (2013). Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. In *Atmospheric Environment* (Vol. 69, pp. 94–108). https://doi.org/10.1016/j.atmosenv.2012.11.009

Bell, M. L., Samet, J. M., & Dominici, F. (2004). Time-series studies of particulate matter. *Annual Review of Public Health*, *25*, 247–280. https://doi.org/10.1146/annurev.publhealth.25.102802.124329

Benkendorf, D. J., & Hawkins, C. P. (2020). Effects of sample size and network depth on a deep learning approach to species distribution modeling. *Ecological Informatics*, *60*(February). https://doi.org/10.1016/j.ecoinf.2020.101137

Borlaza. (2021). Disparities in particulate matter (PM10) origins and oxidative potential at a city scale (Grenoble, France) - Part 2: Sources of PM10 oxidative potential using multiple linear regression analysis and the predictive applicability of multilayer perceptron n. *Atmospheric Chemistry and Physics*, *21*(12), 9719–9739. https://doi.org/10.5194/acp-21-9719-2021

Borlaza, L. J. S., Weber, S., Uzu, G., Jacob, V., Cañete, T., Micallef, S., Trébuchon, C., Slama, R., Favez, O., & Jaffrezo, J. L. (2021). Disparities in particulate matter (PM10) origins and oxidative potential at a city scale (Grenoble, France) - Part 1: Source apportionment at three neighbouring sites. *Atmospheric Chemistry and Physics*, *21*(7), 5415–5437. https://doi.org/10.5194/acp-21-5415-2021

Bourlard, H., & Wellekens, C. J. (1989). *Speech pattern discrimination and multilayer perceptrons*.

Breiman, L. (2001). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *Machine Learning, 12343 LNCS*, 503–515. https://doi.org/10.1007/978-3-030-62008-0_35

Brown, S. G., Eberly, S., Paatero, P., & Norris, G. A. (2015). Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. *Science of the Total Environment*, *518–519*, 626–635. https://doi.org/10.1016/j.scitotenv.2015.01.022

Calas, A., Uzu, G., Besombes, J. L., Martins, J. M. F., Redaelli, M., Weber, S., Charron, A., Albinet, A., Chevrier, F., Brulfert, G., Mesbah, B., Favez, O., & Jaffrezo, J. L. (2019). Seasonal variations and chemical predictors of oxidative potential (OP) of particulate matter (PM), for seven urban French sites. *Atmosphere*, *10*(11). https://doi.org/10.3390/atmos10110698

Calas, A., Uzu, G., Kelly, F. J., Houdier, S., Martins, J. M. F., Thomas, F., Molton, F., Charron, A., Dunster, C., Oliete, A., Jacob, V., Besombes, J. L., Chevrier, F., & Jaffrezo, J. L. (2018). Comparison between five acellular oxidative potential measurement assays performed with detailed chemistry on PM10 samples from the city of Chamonix (France). *Atmospheric Chemistry and Physics*, *18*(11), 7863–7875. https://doi.org/10.5194/acp-18-7863-2018

Calas, A., Uzu, G., Martins, J. M. F., Voisin, Di., Spadini, L., Lacroix, T., & Jaffrezo, J. L. (2017). The importance of simulated lung fluid (SLF) extractions for a more relevant evaluation of the oxidative potential of particulate matter. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/s41598-017-11979-3

Chianese, E., Camastra, F., & Ciaramella, A. (2018). *Spatio-temporal learning in predicting ambient particulate matter concentration by multi-layer perceptron Spatio-temporal Learning in Predicting Ambient Particulate Matter Concentration by Multi-Layer*. *December*. https://doi.org/10.1016/j.ecoinf.2018.12.001

Cho, A., Sioutas, C., Miguel, A. H., Kumagai, Y., Schmitz, D. A., Singh, M., Eiguren-Fernandez, A., & Froines, J. R. (2005). Redox activity of airborne particulate matter at different sites in the Los Angeles Basin. *Environmental Research*, *99*(1), 40–47. https://doi.org/10.1016/j.envres.2005.01.003

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403. https://doi.org/10.1081/QEN-120001878

Crobeddu, B., Aragao-Santiago, L., Bui, L. C., Boland, S., & Baeza Squiban, A. (2017). Oxidative potential of particulate matter 2.5 as predictive indicator of cellular stress. *Environmental Pollution*, *230*, 125–133. https://doi.org/10.1016/j.envpol.2017.06.051

Crouse, D. L., Peters, P. A., Hystad, P., Brook, J. R., van Donkelaar, A., Martin, R. V., Villeneuve, P. J., Jerrett, M., Goldberg, M. S., Arden Pope, C., Brauer, M., Brook, R. D., Robichaud, A., Menard, R., & Burnett, R. T. (2015). Ambient PM2.5, O3, and NO2 exposures and associations with mortality over 16 years of follow-up in the canadian census health and environment cohort (CanCHEC). *Environmental Health Perspectives*, *123*(11), 1180–1186. https://doi.org/10.1289/ehp.1409276

Crouse, D. L., Peters, P. A., van Donkelaar, A., Goldberg, M. S., Villeneuve, P. J., Brion, O., Khan, S., Atari, D. O., Jerrett, M., Pope, C. A., Brauer, M., Brook, J. R., Martin, R. V., Stieb, D., & Burnett, R. T. (2012). Risk of nonaccidental and cardiovascular mortality in relation to long-term exposure to low concentrations of fine particulate matter: A canadian national-level cohort study. *Environmental Health Perspectives*, *120*(5), 708–714. https://doi.org/10.1289/ehp.1104049

Daellenbach, K. R., Uzu, G., Jiang, J., Cassagnes, L.-E., Leni, Z., Vlachou, A., Stefenelli, G., Canonaco, F., Weber, S., Segers, A., & Sources, al. (2020). Sources of particulate-matter air pollution and its oxidative potential in Europe of particulate-matter air pollution and its oxidative potential in Europe. *Nature*, *587*(7834). https://doi.org/10.1038/s41586-020-2902-8ï

Deng, M., Chen, D., Zhang, G., & Cheng, H. (2022). Policy-driven variations in oxidation potential and source apportionment of PM2.5 in Wuhan, central China. *Science of the Total Environment*, *853*(May), 158255. https://doi.org/10.1016/j.scitotenv.2022.158255

Dominici, F. (2004). Time-series analysis of air pollution and mortality: a statistical review. *Research Report (Health Effects Institute)*, *123*.

Dominutti, P. A., Borlaza, L. J. S., Sauvain, J. J., Ngoc Thuy, V. D., Houdier, S., Suarez, G., Jaffrezo, J. L., Tobin, S., Trébuchon, C., Socquet, S., Moussu, E., Mary, G., & Uzu, G. (2023). Source apportionment of oxidative potential depends on the choice of the assay: insights into 5 protocols comparison and implications for mitigation measures. *Environmental Science: Atmospheres*. https://doi.org/10.1039/d3ea00007a

Elangasinghe, M. A., Singhal, N., Dirks, K. N., & Salmond, J. A. (2014). Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmospheric Pollution Research*, *5*(4), 696–708. https://doi.org/10.5094/APR.2014.079

Fadel, M., Courcot, D., Delmaire, G., Roussel, G., Afif, C., & Ledoux, F. (2023). Source apportionment of PM2.5 oxidative potential in an East Mediterranean site. *Science of the Total Environment*, *900*(July). https://doi.org/10.1016/j.scitotenv.2023.165843

Fang, T., Verma, V., T Bates, J., Abrams, J., Klein, M., Strickland, J. M., Sarnat, E. S., Chang, H. H., Mulholland, A. J., Tolbert, E. P., Russell, G. A., & Weber, J. R. (2016). Oxidative potential of ambient water-soluble PM2.5 in the southeastern United States: Contrasts in sources and health associations between ascorbic acid (AA) and dithiothreitol (DTT) assays. *Atmospheric Chemistry and Physics*, *16*(6), 3865–3879. https://doi.org/10.5194/acp-16-3865-2016

Favez, O. (2017). *Traitement harmonisé de jeux de données multi-sites pour l'étude des sources de PM par Positive Matrix Factorization*.

Godri, K. J., Harrison, R. M., Evans, T., Baker, T., Dunster, C., Mudway, I. S., & Kelly, F. J. (2011). Increased oxidative burden associated with traffic component of ambient particulate matter at roadside and Urban background schools sites in London. *PLoS ONE*, *6*(7). https://doi.org/10.1371/journal.pone.0021961

Goldfeld, S. M., & Quandt, R. E. (1965). Some Tests for Homoscedasticity Author (s): Stephen M . Goldfeld and Richard E . Quandt Source : Journal of the American Statistical Association , Jun ., 1965 , Vol . 60 , No . 310 Published by : Taylor & Francis , Ltd . on behalf of the American Statis. *Journal of the American Statistical Association*, *60*(310), 539–547.

Harrell. (2016). Regression Modeling Strategies. *Technometrics*, *45*(2), 170–170. https://doi.org/10.1198/tech.2003.s158

Hastie, T. et. all. (2009). Springer Series in Statistics The Elements of Statistical Learning. *The Mathematical Intelligencer*, *27*(2), 83–85. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *29*(5), 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, *12*(1), 69. https://doi.org/10.2307/1267352

in 't Veld, M., Pandolfi, M., Amato, F., Pérez, N., Reche, C., Dominutti, P., Jaffrezo, J., Alastuey, A., Querol, X., & Uzu, G. (2023a). Discovering oxidative potential (OP) drivers of atmospheric PM10, PM2.5, and PM1 simultaneously in North-Eastern Spain. *Science of the Total Environment*, *857*(August 2022). https://doi.org/10.1016/j.scitotenv.2022.159386

Janssen, N. A. H., Yang, A., Strak, M., Steenhof, M., Hellack, B., Gerlofs-Nijland, M. E., Kuhlbusch, T., Kelly, F., Harrison, R., Brunekreef, B., Hoek, G., & Cassee, F. (2014). Oxidative potential of particulate matter collected at sites with different source characteristics. *Science of the Total Environment*, *472*, 572–581. https://doi.org/10.1016/j.scitotenv.2013.11.099

Kelly, F. J., & Mudway, I. S. (2003). Protein oxidation at the air-lung interface. *Amino Acids*, *25*(3–4), 375–396. https://doi.org/10.1007/s00726-003-0024-x

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Leni, Z., Cassagnes, L. E., Daellenbach, K. R., Haddad, I. El, Vlachou, A., Uzu, G., Prévôt, A. S. H., Jaffrezo, J. L., Baumlin, N., Salathe, M., Baltensperger, U., Dommen, J., & Geiser, M. (2020). Oxidative stress-induced inflammation in susceptible airways by anthropogenic aerosol. *PLoS ONE*, *15*(11 November).

https://doi.org/10.1371/journal.pone.0233425

Li, J., Zhao, S., Xiao, S., Li, X., Wu, S., Zhang, J., & Schwab, J. J. (2023). Source apportionment of water-soluble oxidative potential of PM 2 . 5 in a port city of Xiamen , Southeast China. *Atmospheric Environment*, *314*(June), 120122. https://doi.org/10.1016/j.atmosenv.2023.120122

Li, Xia, T., & Nel, A. E. (2008). The role of oxidative stress in ambient particulate matter-induced lung diseases and its implications in the toxicity of engineered nanoparticles. *Free Radical Biology and Medicine*, *44*(9), 1689–1699. https://doi.org/10.1016/j.freeradbiomed.2008.01.028

Liu, & Ng. (2023). *Toxicity of Atmospheric Aerosols: Methodologies & Assays*.

Liu, W. J., Xu, Y. S., Liu, W. X., Liu, Q. Y., Yu, S. Y., Liu, Y., Wang, X., & Tao, S. (2018). Oxidative potential of ambient PM2.5 in the coastal cities of the Bohai Sea, northern China: Seasonal variation and source apportionment. *Environmental Pollution*, *236*, 514–528. https://doi.org/10.1016/j.envpol.2018.01.116

Lodovici, M., & Bigagli, E. (2011). Oxidative stress and air pollution exposure. *Journal of Toxicology*, *2011*. https://doi.org/10.1155/2011/487074

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*, *20*(1), 20–33. https://doi.org/10.1080/10888438.2015.1107073

McCullagh. (1989). Generalized linear models. In *Statistical Models in S* (pp. 195–247). https://doi.org/10.1201/9780203738535

Mudway, I. S., Kelly, F. J., & Holgate, S. T. (2020). Oxidative stress in air pollution research. In *Free Radical Biology and Medicine* (Vol. 151, pp. 2–6). Elsevier Inc. https://doi.org/10.1016/j.freeradbiomed.2020.04.031

Nelin, T. D., Joseph, A. M., Gorr, M. W., & Wold, L. E. (2012). Direct and indirect effects of particulate matter on the cardiovascular system. *Toxicology Letters*, *208*(3), 293–299. https://doi.org/10.1016/j.toxlet.2011.11.008

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Paatero, P., & Hopke, P. K. (2009). Rotational tools for factor analytic models. *Journal of Chemometrics*, *23*(2), 91–100. https://doi.org/10.1002/cem.1197

Paatero, P., & Tappert, U. (1994). POSITIVE MATRIX FACTORIZATION: A NON-NEGATIVE FACTOR MODEL WITH OPTIMAL UTILIZATION OF ERROR ESTIMATES OF DATA VALUES*. In *ENVIRONMETRICS* (Vol. 5).

Pearce, J., & Ferrier, S. (2000). An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, *128*(2–3), 127–147. https://doi.org/10.1016/S0304-3800(99)00227-6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pelucchi, C., Negri, E., Gallus, S., Boffetta, P., Tramacere, I., & La Vecchia, C. (2009). Long-term particulate matter exposure and mortality: A review of European epidemiological studies. *BMC Public Health*, *9*, 1–8. https://doi.org/10.1186/1471-2458-9-453

Peng, R. D., Bell, M. L., Geyh, A. S., McDermott, A., Zeger, S. L., Samet, J. M., & Dominici, F. (2009). Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental Health Perspectives*, *117*(6), 957–963. https://doi.org/10.1289/ehp.0800185

Pietrogrande, M. C., Romanato, L., & Russo, M. (2022). Synergistic and Antagonistic Effects of Aerosol Components on Its Oxidative Potential as Predictor of Particle Toxicity. *Toxics*, *10*(4). https://doi.org/10.3390/toxics10040196

Pope, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air and Waste Management Association*, *56*(6), 709–742. https://doi.org/10.1080/10473289.2006.10464485

Rao, X., Zhong, J., Brook, R. D., & Rajagopalan, S. (2018). Effect of Particulate Matter Air Pollution on Cardiovascular Oxidative Stress Pathways. *Antioxidants and Redox Signaling*, *28*(9), 797–818. https://doi.org/10.1089/ars.2017.7394

Raudys, S. J., & Jain, A. K. (1991). Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vol. 13, Issue 3, pp. 252–264). https://doi.org/10.1109/34.75512

Rosenblad, A. (2011). The Concise Encyclopedia of Statistics. In *Journal of Applied Statistics* (Vol. 38, Issue 4). https://doi.org/10.1080/02664760903075614

Samake, A., Uzu, G., Martins, J. M. F., Calas, A., Vince, E., Parat, S., & Jaffrezo, J. L. (2017). The unexpected role of bioaerosols in the Oxidative Potential of PM. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-11178-0

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Serafeim, E., Besis, A., Kouras, A., Farias, C. N., Yera, A. B., Pereira, G. M., Samara, C., & de Castro Vasconcellos, P. (2023). Oxidative potential of ambient PM2.5 from São Paulo, Brazil: Variations, associations with chemical components and source apportionment. *Atmospheric Environment*, *298*. https://doi.org/10.1016/j.atmosenv.2023.119593

Shangguan, Y., Zhuang, X., Querol, X., Li, B., Moreno, N., Trechera, P., Sola, P. C., Uzu, G., & Li, J. (2022). Characterization of deposited dust and its respirable fractions in underground coal mines: Implications for oxidative potential-driving species and source apportionment. *International Journal of Coal Geology*, *258*(December 2021). https://doi.org/10.1016/j.coal.2022.104017

Stevanović, S., Jovanović, M. V., Jovašević-Stojanović, M. V., & Ristovski, Z. (2023). SOURCE APPORTIONMENT OF OXIDATIVE POTENTIAL What We Know So Far. *Thermal Science*, *27*(3), 2347–2357. https://doi.org/10.2298/TSCI221107111S

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*(1), 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Szigeti, T., Dunster, C., Cattaneo, A., Cavallo, D., Spinazzè, A., Saraga, D. E., Sakellaris, I. A., de Kluizenaar, Y., Cornelissen, E. J. M., Hänninen, O., Peltonen, M., Calzolai, G., Lucarelli, F., Mandin, C., Bartzis, J. G., Záray, G., & Kelly, F. J. (2016). Oxidative potential and chemical composition of PM2.5 in office buildings across Europe - The OFFICAIR study. *Environment International*, *92–93*, 324–333. https://doi.org/10.1016/j.envint.2016.04.015

Szigeti, T., Óvári, M., Dunster, C., Kelly, F. J., Lucarelli, F., & Záray, G. (2015). Changes in chemical composition and oxidative potential of urban PM2.5 between 2010 and 2013 in Hungary. *Science of the Total Environment*, *518–519*, 534–544. https://doi.org/10.1016/j.scitotenv.2015.03.025

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Verma, V., Fang, T., Guo, H., King, L., Bates, J. T., Peltier, R. E., Edgerton, E., Russell, A. G., & Weber, R. J. (2014). Reactive oxygen species associated with water-soluble PM2.5 in the southeastern United States: Spatiotemporal trends and source apportionment. *Atmospheric Chemistry and Physics*, *14*(23), 12915–12930. https://doi.org/10.5194/acp-14-12915-2014

Viana, M., Kuhlbusch, T. A. J., Querol, X., Alastuey, A., Harrison, R. M., Hopke, P. K., Winiwarter, W., Vallius, M., Szidat, S., Prévôt, A. S. H., Hueglin, C., Bloemen, H., Wåhlin, P., Vecchi, R., Miranda, A. I., Kasper-Giebl, A., Maenhaut, W., & Hitzenberger, R. (2008). Source apportionment of particulate matter in Europe: A review of methods and results. In *Journal of Aerosol Science* (Vol. 39, Issue 10, pp. 827–849). Elsevier Ltd. https://doi.org/10.1016/j.jaerosci.2008.05.007

Vida, M., Foret, G., Siour, G., Coman, A., Weber, S., Favez, O., Jaffrezo, J., Pontet, S., Mesbah, B., Gille, G., Zhang, S., Chevrier, F., Pallares, C., Uzu, G., & Beekmann, M. (2024). Oxidative potential modelling of

PM10: a 2-year study over France. To be summited to *ACDP*.

Wang, D., Yang, X., Lu, H., Li, D., Xu, H., Luo, Y., Sun, J., Hang Ho, S. S., & Shen, Z. (2023). Oxidative potential of atmospheric brown carbon in six Chinese megacities: Seasonal variation and source apportionment. *Atmospheric Environment*, *309*(March), 119909. https://doi.org/10.1016/j.atmosenv.2023.119909

Wang, Wang, M., Li, S., Sun, H., Mu, Z., Zhang, L., Li, Y., & Chen, Q. (2020). Study on the oxidation potential of the water-soluble components of ambient PM2.5 over Xi'an, China: Pollution levels, source apportionment and transport pathways. *Environment International*, *136*(January), 105515. https://doi.org/10.1016/j.envint.2020.105515

Weber, S., Salameh, D., Albinet, A., Alleman, L. Y., Waked, A., Besombes, J. L., Jacob, V., Guillaud, G., Meshbah, B., Rocq, B., Hulin, A., Dominik-Sègue, M., Chrétien, E., Jaffrezo, J. L., & Favez, O. (2019). Comparison of PM10 sources profiles at 15 french sites using a harmonized constrained positive matrix factorization approach. *Atmosphere*, *10*(6). https://doi.org/10.3390/atmos10060310

Weber, S., Uzu, G., Calas, A., Chevrier, F., Besombes, J. L., Charron, A., Salameh, D., Ježek, I., Moĉnik, G., & Jaffrezo, J. L. (2018). An apportionment method for the oxidative potential of atmospheric particulate matter sources: Application to a one-year study in Chamonix, France. *Atmospheric Chemistry and Physics*, *18*(13), 9617–9629. https://doi.org/10.5194/acp-18-9617-2018

Weber, S., Uzu, G., Favez, O., Borlaza, L. J. S., Calas, A., Salameh, D., Chevrier, F., Allard, J., Besombes, J. L., Albinet, A., Pontet, S., Mesbah, B., Gille, G., Zhang, S., Pallares, C., Leoz-Garziandia, E., & Jaffrezo, J. L. (2021). Source apportionment of atmospheric PM10 oxidative potential: Synthesis of 15 year-round urban datasets in France. *Atmospheric Chemistry and Physics*, *21*(14), 11353–11378. https://doi.org/10.5194/acp-21-11353-2021

WHO. (2021). *WHO global air quality guidelines*.

Williams, M., Gomez Grajales, C. A., & Kurkiewicz, D. (2013). Assumptions of Multiple Regression: Correcting Two Misconceptions - Practical Assessment, Research & Evaluation. *Practical Assessment, Research, and Evaluation (PARE)*, *18*(11), 1–16. https://scholarworks.umass.edu/pare/vol18/iss1/11

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., Elith, J., Dudík, M., Ferrier, S., Huettmann, F., Leathwick, J. R., Lehmann, A., Lohmann, L., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. C., … Zimmermann, N. E. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Yang, A., Jedynska, A., Hellack, B., Kooter, I., Hoek, G., Brunekreef, B., Kuhlbusch, T. A. J., Cassee, F. R., & Janssen, N. A. H. (2014). Measurement of the oxidative potential of PM2.5 and its constituents: The effect of extraction solvent and filter type. *Atmospheric Environment*, *83*, 35–42. https://doi.org/10.1016/j.atmosenv.2013.10.049

Yu, Guo, S., Xu, R., Ye, T., Li, S., Sim, M. R., Abramson, M. J., & Guo, Y. (2021). Cohort studies of long-term exposure to outdoor particulate matter and risks of cancer: A systematic review and meta-analysis. *Innovation*, *2*(3), 100143. https://doi.org/10.1016/j.xinn.2021.100143

Yu, S. Y., Liu, W. J., Xu, Y. S., Yi, K., Zhou, M., Tao, S., & Liu, W. X. (2019). Characteristics and oxidative potential of atmospheric PM2.5 in Beijing: Source apportionment and seasonal variation. *Science of the Total Environment*, *650*, 277–287. https://doi.org/10.1016/j.scitotenv.2018.09.021

Zhang, Y., Albinet, A., Petit, J. E., Jacob, V., Chevrier, F., Gille, G., Pontet, S., Chrétien, E., Dominik-Sègue, M., Levigoureux, G., Močnik, G., Gros, V., Jaffrezo, J. L., & Favez, O. (2020). Substantial brown carbon emissions from wintertime residential wood burning over France. *Science of the Total Environment*, *743*. https://doi.org/10.1016/j.scitotenv.2020.140752